DEDUP : processus de dédoublonnage des notices

Primo offre un mécanisme permettant de fusionner des notices entre elles. Ce processus intervient à la fin du pipe pour toutes les ressources moissonnées localement. Il est ainsi possible de fusionner une notice issue du répertoire Unimarc Alma avec une autre issue du répertoire Marc21 Alma.

Il est en revanche impossible de fusionner une notice importée localement (par exemple issue d'Alma) avec une notice issue de Primo central

Une notice fusionnée dans Primo porte normalement deux mentions de disponibilités.

Son identifiant est préfixé par **dedupmrg**.

Fonctionnement du DEDUP

Au moment de la normalisation le système créé différentes clefs qui vont alimenter l'algorithme de fusion. Ces clefs sont définies au niveau des champs dedup de la notice pnx. Le processus de fusion se déroule en 3 étapes. A chacune de ces étapes, il va utiliser des clefs de types différents.

Choix des règles de matching à appliquer en fonction du type de document : clef de type t

La clef de type T permet d'indiquer à l'algorithme quelle famille de règles appliquer en fonction du type de document. Il en existe 3 :

- T=1 pour la règle qui s'applique pour les notices de documents hors publication en série
- T=2 pour la règle qui s'applique pour les publications en série
- T=3 pour la règle qui s'applique pour les articles

Pour exclure une notice de la fusion, il suffit d'appliquer la **valeur 99 à la clef T**

Choix des candidats doublon : clef de type C

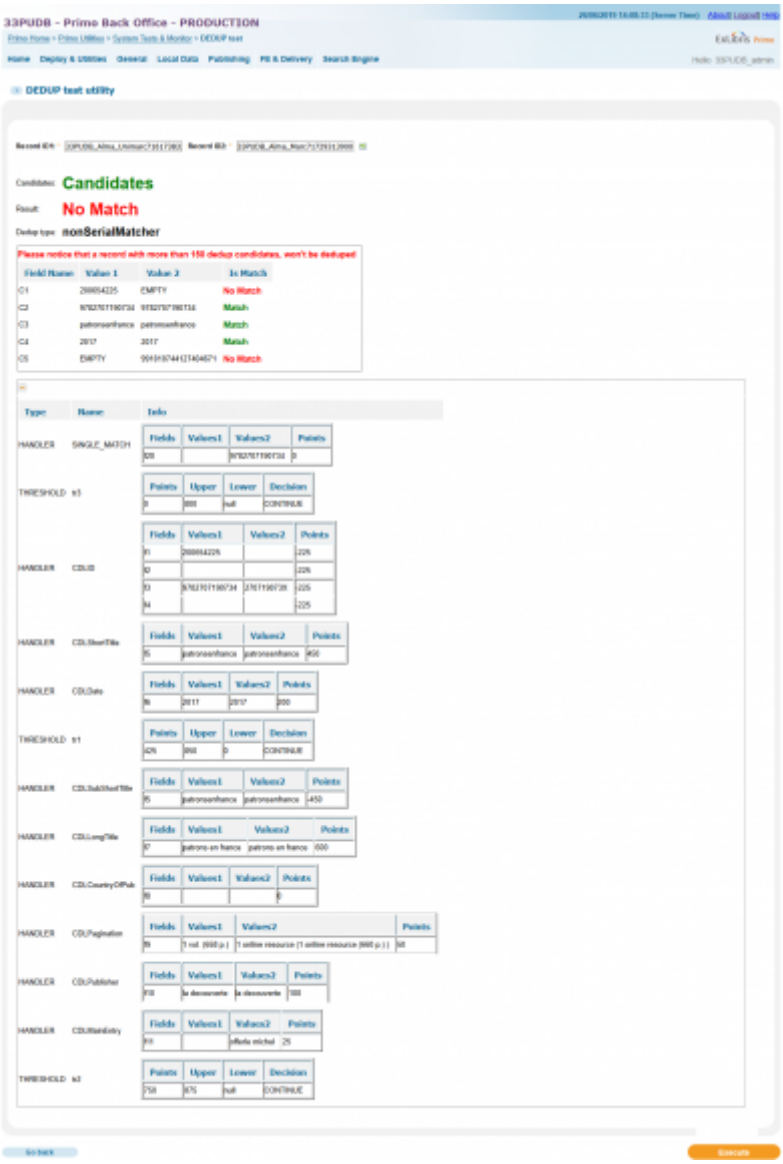
A ce stade, le système utilise les clefs de type C pour repérer les doublons potentiels. Ces clefs varient en fonction du type de règles appliquées. On y retrouve, le titres, les identifiants normalisés, les dates de publications, …

Phase de fusion : clef de type F

Le système fait passer une série de tests aux notices candidates en s'appuyant sur les clefs de types F. Chaque test réussi donne des points à la fusion. Certains tests échoués donnent des malus.

Par exemple, pour les périodiques, si il y a une correspondance parfaite du titre des deux notices, le système attribue 600 points. Si la correspondance est partielle, il attribue 175 points. En cas de non correspondance, il donne un malus de -600 points.

Les tests sont ventilés sur 3 étapes. A chaque étape le système additionne le nombre de points obtenus lors de chaque test. A l'issue de la première étape, si la somme des points obtenus par les deux notices est supérieure à 800 points, le système valide la fusion et termine le processus. Si le nombre de points est inférieur le système va jouer la batterie des tests suivants. Pour les deux étapes suivantes la somme des points devra être supérieure à 875 pour valider la fusion.



Les tests sont détaillés dans la documentation d'Exlibris.
Il est possible de tester la fusion des notices grâce au **DEDUP test utility** sous Primo Utilities>System tests & monitor>Dedup Test

__Soulignage_ _

Mise en œuvre du dedup à Bordeaux

Le réseau a décidé de fusionner entre elles les notices décrivant une même expression ayant des manifestations différentes.
Pour l'instant, l'effort a été concentré sur la fusion des notices décrivant un document physique (issue du répertoire Unimarc Alma) avec la notice décrivant le même document sur support électronique (Alma Marc 21).

Amélioration des clefs pour la détection des candidats (C)

Pour 33PUDB_AlmaUnimarc

Champ C1

Le C1 est normalement dédié aux identifiants système. Par défaut il est construit à partir du numéro de la bibliographie nationale (020 $b). Nous avons ajouté le :

- PPN valide et erronné de la notice (035 $a(PPN)* et $z(PPN)*)
- NNT de la notice de thèse (035 $a(NNT)*)
- PPN de la notice de l’édition sur un autre support (452 $1)
- PPN de la notice du document original (455 $1)
- PPN de la notice de la reproduction du document (456 $1)

Pour les zones 4XX on ne créé pas de clef si la zone 200 comporte un $$h (numéro de volume) et si le champs 4## n'en contient pas. Ceci afin d'éviter des fusions non désirées lorsque on a lier une notice décrivant une partie composante à une notice de regroupement.

Champ C2

Le C2 contient les identifiants bibliographiques. La clef est construite à partir de :

- l’ISBN valide ou invalide du document (010 $a ou $z) [MAJ 25/06/2019] transformé en ISBN 13 si ISBN 10
- l’ISSN valide ou invalide du document (011 $a ou $z)
- ISSN ou ISBN de la notice de l’édition sur un autre support (452 $x,y) [MAJ 25/06/2019] transformé en ISBN 13 si ISBN 10
- ISSN ou ISBN de la notice du document original (455 $x,y) [MAJ 25/06/2019] transformé en ISBN 13 si ISBN 10
- ISSN ou ISBN de la notice de la reproduction du document (456 $x,y) [MAJ 25/06/2019] transformé en ISBN 13 si ISBN 10

Pour 33PUDB_AlmaMarc

Champ C1

Le C1 sera à retravailler quand nous aurons à relier nos portfolios à des notices SUDOC Marc21

Champ C2

La clef est construite à partir de :

- l'ISBN valide ou invalide du document (020 $a ou $z) [MAJ 25/06/2019] transformé en ISBN 13 si ISBN 10
- l'ISSN valide ou invalide du document (022 $a ou $z)
- ISSN ou ISBN de la notice de l'édition sur un autre support (776 $x,z) [MAJ 25/06/2019] transformé en ISBN 13 si ISBN 10

Amélioration des clefs pour la réalisation des matchings (f)

Champ f20

La clef f20 est une super clef. C'est le premier test qui est mené et en cas de correspondance il rapporte directement 800 points. La fusion est donc validée et aucun autre test n'est réalisé. ⚠**Le problème c'est qu'elle ne peut pas être multi-valuée** L'idée est de construire cette clef en fonction de la source de la notice traitée et de la source de données avec laquelle on souhaite réalisée la fusion. Je propose de construire la clef selon ces règles.

Source	**Types de document**	**Source candidate au matching**	**Identifiant fort sur lequel réaliser le matching**
————————-	————————————	————————————-	——————————————————————-
33PUDB_1886	Livres Anciens	33PUDB_Alma_Unimarc	PPN de la notice
33PUDB_Alma_Marc	Livres électroniques	33PUDB_Alma_Unimarc	ISBN du document sous un autre format (776$z)
33PUDB_Alma_Marc	Revues électroniques	33PUDB_Alma_Unimarc	ISSN du document sous un autre format (776$x)
33PUDB_Alma_Unimarc	Livres électroniques	33PUDB_Alma_Marc	ISBN du document (010 $a)
33PUDB_Alma_Unimarc	Revues électroniques	33PUDB_Alma_Marc	ISSN du document (011 $a)
33PUDB_Alma_Unimarc	Livres Anciens	33PUDB_1886	PPN de la notice (035$a(PPN)*)
33PUDB_Alma_Unimarc	Livres Anciens	33PUDB_BabordNum	PPN de la notice (035$a(PPN)*)
33PUDB_Alma_Unimarc	Journeaux Anciens	33PUDB_BabordNum	PPN de la notice (035$a(PPN)*)
33PUDB_Alma_Unimarc	Mémoires et thèses d'exercices	33PUDB_DUMAS_UB	NNT (029$a ou 035$a(NNT))
33PUDB_apprentoile	-	-	-
33PUDB_BabordNum	Livres Anciens	33PUDB_ALMA_Unimarc	PPN de la version originale
33PUDB_BabordNum	Journeaux Anciens	33PUDB_ALMA_Unimarc	PPN de la version originale
33PUDB_CanalU	-	-	-
33PUDB_DUMAS_UB	Mémoires et thèses d'exercices	33PUDB_ALMA_Unimarc	NNT (display/identifier)
33PUDB_DNO	-	-	-

33PUDB_ERMS_UB	-	-	-

Effets de bords possibles pour la source Unimarc :
On peut avoir des notices décrivant des œuvres différentes sous un même ISBN. Dans le cas de monographies en plusieurs volumes ou d'un coffret, un établissement peut être localisé sous la notice du coffret alors qu'un autre a préféré se localiser sous chaque volume. Dans ce cas Primo va fusionner la notice du coffret avec un des volumes en fonction de l'ISBN qui aura été utilisé (c'est le premier ISBN qui est sélectionné par défaut).

Je vois deux solutions pour nous prémunir de ces cas de fusions illégitimes :

- **On pourrait jouer sur la présence d'un champ 463 ou 464 pour exclure de la fusion les coffrets ou les monographies en plusieurs volumes. [stratégie ceinture]** (ex : (PPN)123692962).<note important>Mais les notices ne détaillent pas toujours les parties composantes (ex (PPN)115276610).</note>
- **On pourrait limiter la construction de la clef aux seuls notices unimarc signalant une édition sous un autre support (présence d'un champ 452,455 ou 456)[stratégie bretelle]**. <note important>Cela restreint quelques peut les possibilités de fusion. Mais si on détecte une notice candidate à la fusion, il suffit d'ajouter dans le SUDOC une 452. Cela contribuera à l'enrichissement du catalogue collectif</note>

Dans l'immédiat nous avons juste appliquer la stratégie bretelle.
[mAJ 26/02/2021] La stratégie bretelle était trop restrictive on passe à la stratégie ceinture.
[mAJ 05/06/2019] De nombreuses notices de monographies ne fusionnaient pas car les ISBN présents dans les notices marc21 étaient des ISBN 10 et les ISBN des notices candidates à la fusion en Unimarc des ISBN 13. En convertissant systématiquement l'ISBN de 10 à 13 lors de la construction des clefs C2 et F20 nous avons augmenté le nombre de notices fusionnées.

33PUDB_Alma_Unimarc

- Si il s'agit d'une thèse d'exercice numérisée (display/type = dissertation) avec un lien vers dumas (856 contient dumas.ccsd.cnrs.fr) on prent le NNT
- Sinon si un ISBN est présent on prend l'ISBN valide si la notice possède un champ 452 et 455 ou 456 [MAJ 25/06/2019] transformé en ISBN 13 si ISBN 10
- Sinon si un ISSN est présent on prend l'ISSN valide si la notice possède un champ 452 et 455 ou 456

33PUDB_Alma_Marc

- On prend l'ISSN ou l'ISBN de la notice du document sur un autre support (776 $x ou $z) [MAJ 25/06/2019] transformé en ISBN 13 si ISBN 10

Champ F1

33PUDB_Alma_Unimarc

- Numéro de la bibliographie nationale (020 $b)
- PPN validede la notice (035 $a(PPN)*)
- NNT de la notice de thèse (035 $a(NNT)*)
- PPN de la notice de l'édition sur un autre support (452 $1)

- PPN de la notice du document original (455 $1)
- PPN de la notice de la reproduction du document (456 $1)

33PUDB_Alma_Marc

A traiter plus tard

Champ F2

33PUDB_Alma_Unimarc

- Numéro invalide de la bibliographie nationale (020 $z)
- PPN invalide de la notice (035 $z(PPN)*)

33PUDB_Alma_Marc

A traiter plus tard

Champ F3

33PUDB_Alma_Unimarc

- l'ISBN valide du document (010 $a)
- l'ISSN valide du document (011 $a)
- ISSN ou ISBN de la notice de l'édition sur un autre support (452 $x,y)
- ISSN ou ISBN de la notice du document original (455 $x,y)
- ISSN ou ISBN de la notice de la reproduction du document (456 $x,y)

33PUDB_Alma_Marc

- l'ISBN valide du document (020 $a)
- l'ISSN validedu document (022 $a)
- ISSN ou ISBN de la notice de l'édition sur un autre support (776 $x,z)

Champ F4

33PUDB_Alma_Unimarc

- l'ISBN invalide du document (010 $z)
- l'ISSN valide du document (011 $z)

33PUDB_Alma_Marc

- l’ISBN invalide du document (020 $z)
- l’ISSN invalide du document (022 $z)

Vérification à faire lorsque deux notices n'ont pas fusionnées

Lorsque deux notices décrivant deux manifestations d'un même document n'ont pas fusionnées :

1. Assurez-vous qu'un lien vers une notice sur un autre support a bien été établi (452,455 ou 456). Si cela n'est pas le cas l'ajout de ce champ dans la notice du Sudoc permettra la fusion des deux notices.
2. Vérifiez que des identifiants renseignés en 010,011,452,455 ou 456 (pour Unimarc) & 020,021 et 776 (pour Marc21) sont corrects
3. Utilisez l'utilitaire de test

Champs retenus/exclus à l'issue de la fusion

Globalement le système conserve tous les champs de la notice PNX à part pour la section display où il privilégie les données de la notice préférée. L’ordre de préférence est défini dans la table **la table de mapping Preferred Record-Delivery Category Priority** en fonction du mode d'accès aux documents décrits.
Nous avons configuré la table ainsi :

1. Alma-P : document physique en provenance d'Alma
2. Physical Item : document physique d'une autre source
3. Alma-E : document électronique en provenance d'Alma
4. Alma-D : document numérique en provenance d'Alma
5. Online Resource : document électronique en provenance d'une autre source
6. SFX Resource
7. Metalib Resource
8. Microform

D'après le support Exlibris Primo va ussi privilégier la notice qui a le plus de champs locaux. Si les paramètres de la table semblent ne pas être respectés ajoutés via les règles de normalisation des champs locaux aux notices pour lesquelles vous voulez conserver les données à l'affichage.

Doublement des champs locaux

Nous utilisons des champ locaux pour personnaliser l'affichage du brief display. Ces champs locaux sont ajoutés pour les notices unimarc comme les notices marc 21. Or, lors de la fusion des notices Primo conserve la totalité des champs locaux. Nous avons dû développer un module angular pour ne conserver qu'un seul champ local.

From:
https://rebub-sgbm.frama.wiki/ - **Le wiki du SGBm du Rebub**

Permanent link:
https://rebub-sgbm.frama.wiki/new:administration_configuration:primo:dedupfrbr:dedup?rev=1614338678

Last update: **2021/02/26 12:24**